Final Progress Report for NASA Grant NNG06GH15G

Principal Investigator:

Robert J. Brunner
Department of Astronomy & NCSA
University of Illinois
1002 West Green Street
Urbana, IL 61853
(217) 244-6099

Summary:

This proposal, which was funded from March 1, 2006 through February 28, 2010 (note this includes a one year NCE), has successfully exceeded all milestones. This work was initiated to address the challenge astronomers, and specifically cosmologists now face in extracting scientific answers from the mountains of data now being generated in our field, including those from NASA missions. Specifically, we proposed to develop advanced astrophysical algorithms that we be deployed on novel supercomputing hardware to (a) robustly classify sources from terascale datasets; and (b) calculate cosmological statistics of the resulting datasets, including n-point correlation measurements. During the period of this award, we have performed a robust classification of the entire SDSS data set, while also determining photometric redshift probability distribution functions for these same sources. We have developed and deployed a fast correlation code on the NCSA supercomputing fabric. We have extended this code to operate on a variety of novel supercomputing technologies, including FPGA, GPU, and Cell based systems, achieving significant speed-ups in the resultant algorithms. Given these initial successes, we extended our work to include algorithms for quantifying photometric and spectral variability, a subject which is becoming increasingly more important with funded projects such as PAN-STARRS and LSST. Finally, we have used our expertise to help others leverage supercomputing technology to tackle their own scientific challenges in a variety of different scientific domains.

Review

Our initial efforts focused on the application of machine learning algorithms to the large-scale, robust classification of sources in Terascale data sets. Our first work was the classification of over 140 million sources from the third data release of the Sloan Digital Sky Survey (SDSS) into three classes: Stars, Galaxies, neither Star nor Galaxy, using Decision Trees (Ball et al. 2006). By applying a blind comparison with deeper test data, we demonstrated (see Figure 1) that our classifications were overall robust and sufficiently faint to enable follow-on scientific analysis of the generated data. We followed this initial work, by extending our approach to (a) utilize a new class of learning algorithms, namely instance-based learning, (b) incorporate a new source category for AGNs (Ball et al. 2007), and (c) to also calculate photometric redshifts for sources

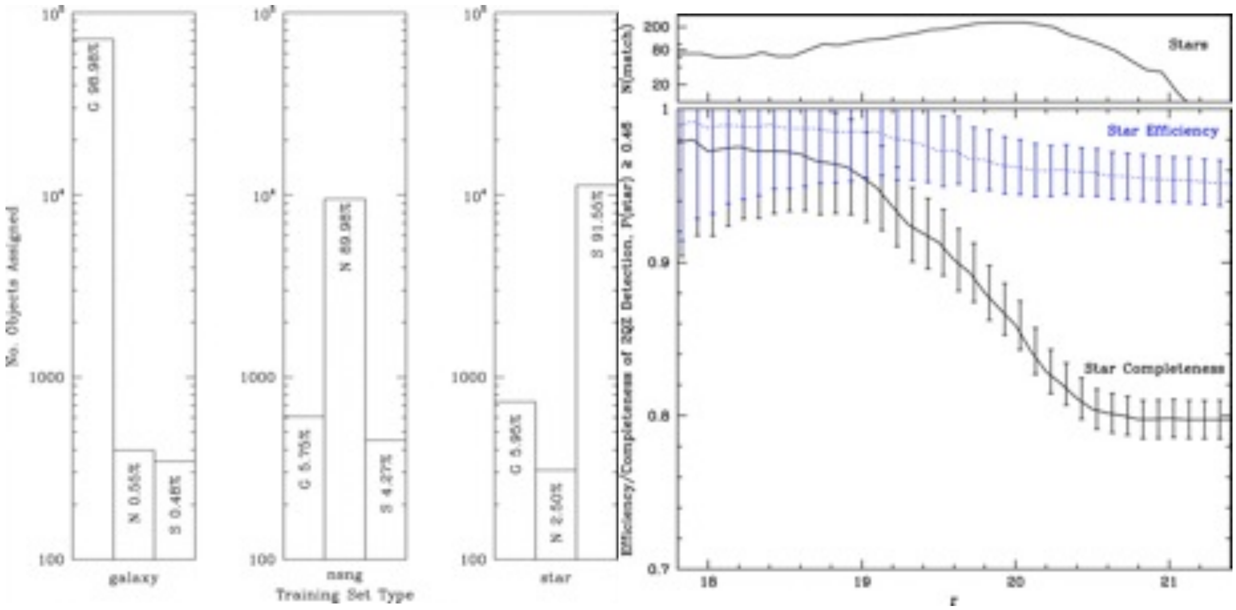

Figure 1: Accuracy of our SDSS DR3 classification work. (Left) Results from a comparison with the testing data. (Right) Results from a blind comparison (Ball et al. 2006).

(Ball et al. 2008). This last item leveraged a new technique we developed that allowed the machine learning algorithm to sample the uncertainty in an objects measured parameters to accurately model the probability distribution function for the source's photometric redshift (Figure 2), which greatly enhances the efficacy of the resulting redshift estimate (see, e.g., Myers et al. 2010).

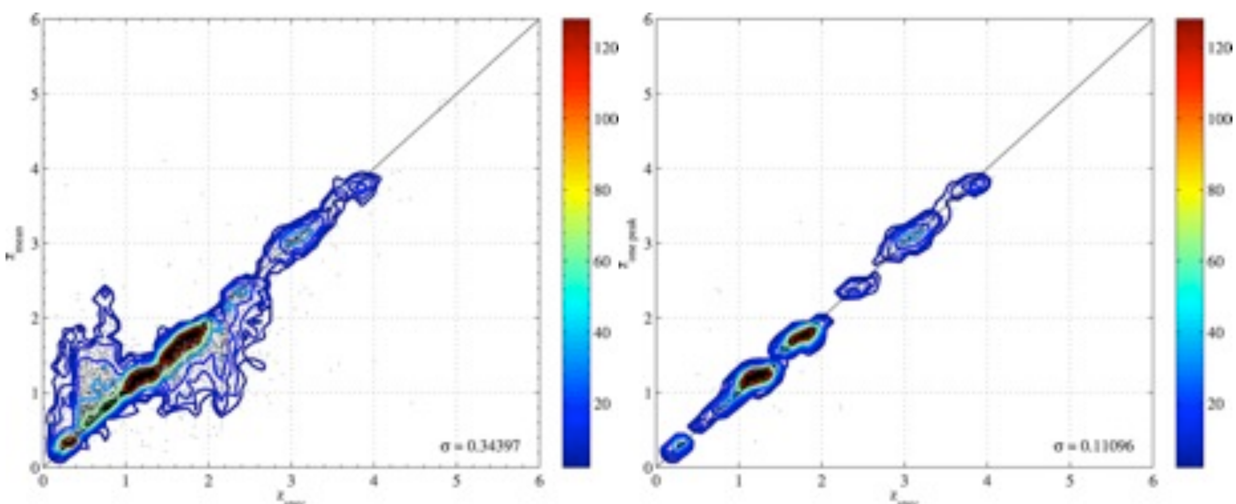

Figure 2: The calculation of photometric redshifts for quasars in the SDSS. (Left) Our initial machine learning results, which are significantly better than previously published efforts. (Right) Our results when we incorporate information from the photometric redshift probability distribution function, achieving a better than a factor of three improvement (Ball et al. 2008).

All of the computations used in these works were performed on the supercomputing resources at NCSA. This required developing new computational frameworks to facilitate the data movement, processing and statistical characterization (see, e.g., Ball and Brunner 2010).

The next effort focused on developing and deploying advanced clustering measurement codes. Our first effort was the development of a software package, written in python, to calculate auto-correlation functions for large data sets across a massively distributed system. This code was first utilized in the auto-correlation measurement of photometrically classified quasars (Myers et a. 2006), which novelly demonstrated the efficacy of photometrically classified samples for cosmological analyses. Subsequent refinements of the code and samples provided further insight into the relationship between quasars, their host galaxies, and their parent dark matter halos (Myers et al. 2007a, 2007b). Our next effort focused on modifying this code to support n-point pixel-based auto-correlation and cross-correlation measurements (Figure 3) for millions of

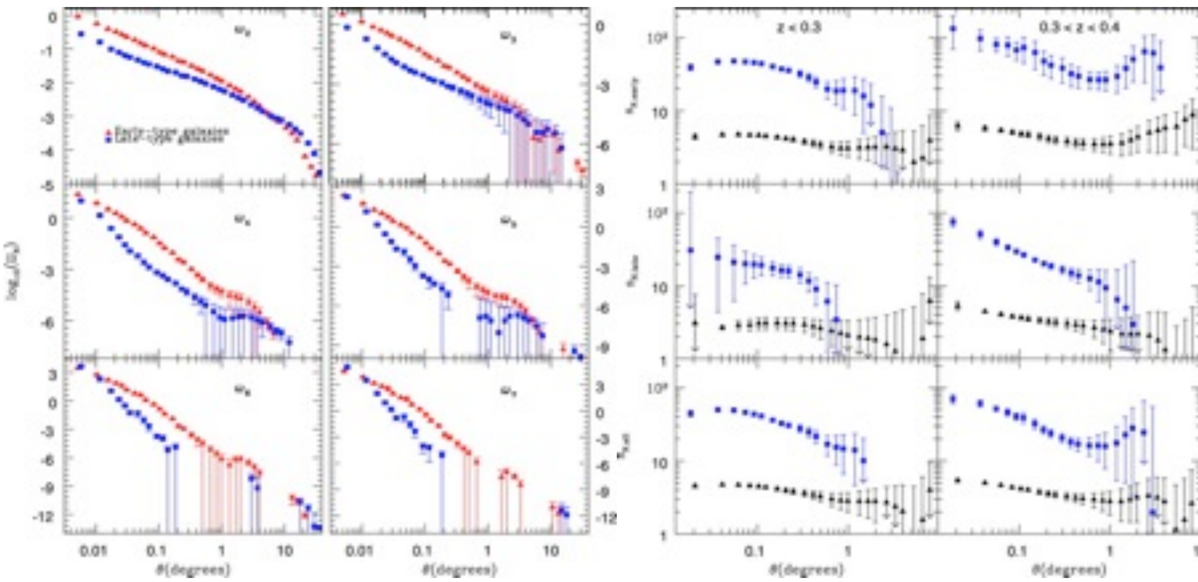

Figure 3: N-Point correlation measurements from tens of millions of SDSS galaxies (Ross et al. 2007). (Left) calculation of photometric redshifts for quasars in the SDSS. (Left) Angular n-point correlation measurements as a function of galaxy type (using photometric classifications). (Right) Projected spatial n-point correlations as a function of redshift (assigned photometrically).

galaxies. With this new code, we were able to make the most precise measurements of galaxy higher order correlation functions (Ross et al. 2006, 2007, 2008).

We next worked to generalize this code to support (Figure 4) point-based auto-correlation and cross-correlation measurements for millions of galaxies (Ross et al. 2009) and cross-correlation measurements of galaxies and quasar absorption systems (Lundgren et al. 2009). The new refinements were demonstrated in operation on Teragrid condor and traditional cluster-based supercomputing resources. While important, these efforts indicated that full point-based correlation measurements for next generation data sets would require a different approach. As a result, we developed a fully distributed hybrid MPI/OpenMP point-based correlation code that has been successfully tested and deployed on Teragrid hybrid HPG systems (Dolence and Brunner 2008). This new code has been successfully run on data sets containing tens of millions of galaxies, calculating fully jack-knifed auto-correlations out to angular scales of tens of degrees. This code leverages MPI to spread jobs across multiple nodes, and uses OpenMP to

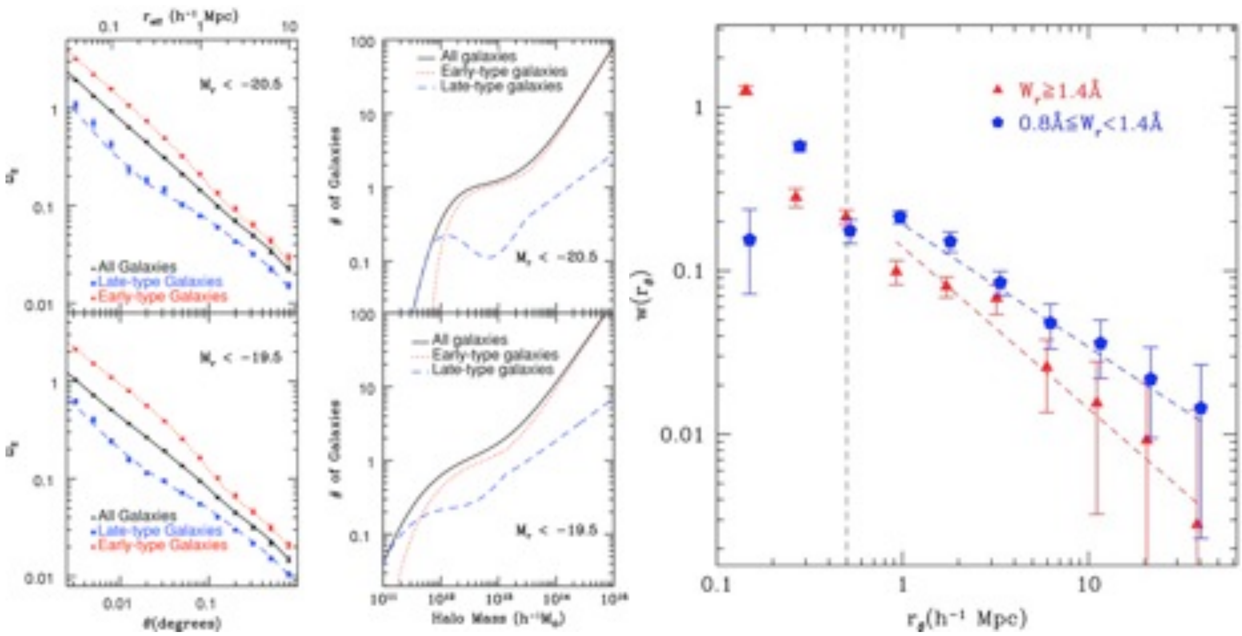

Figure 3:Correlation measurements using our technology on large SDSS data sets. (Left) Several million galaxies auto-correlated by galaxy type (using photometric classifications) with our novel Halo Occupation Distribution modeling (Ross et al. 2010). (Right) Quasar MgII absorptions systems cross-correlated with SDSS galaxies (Lundgren et al. 2009).

spread computations across cores within an individual node. In this manner, we were able to achieve both wide and deep parallelism across these traditional computational systems.

However, in order to continue to accelerate calculations, computational systems are starting to leverage non-traditional technologies including Field Programmable Gate Arrays, Cell-based systems (used in set-top game systems) and Graphical Processing Units. As a result, a large part of the research carried out under this award explored the applicability of these systems to astrophysical algorithms. The primary algorithm used was the two-point angular correlation function (TPACF), a brute-force version of which was ported to several different reconfigurable computational systems (Kindratenko et al. 2007a, 2007b, 2009; Brunner et al. 2007), where we saw, on average, speed-ups of several tens over the same algorithm on a single core system.

As the cost of these systems remains prohibitively high as compared to the commodity graphical processing systems, we transitioned our research over to the many-core systems leveraging nVidia graphical processing units found in the traditional desktop computers. Our implementation leveraging GPUs produced between 50x to over 200x speed-ups relative to a single core desktop, depending on the GPU board used and whether the calculation involved single- or double-precision calculations (Roeh et al. 2009). Interestingly enough, porting the algorithm to these novel architectures demonstrated different performance bottlenecks (such as bin accumulations, where due to non-atomic addition operations in the version of nVidia CUDA we were using, we had to manage these events more carefully. Future versions of CUDA or openCL will likely remove this bottlenecks, likely increasing the efficacy of these algorithms on non-traditional architectures.

We also ported several other algorithms, including an n-nearest neighbor classification algorithm, a pixel-based angular power-spectrum estimator, and a Spherical Harmonic Transform (the last two, along with our first version of the TPACF, served as projects in the ECE498AL course run by Professor Wen-mei Hu and David Kirk), which is used for calculating fast angular power spectra for CMB maps. In all cases, speed-ups were observed, but the nascent software tools limited the global efficacy of these codes. Given the continued industry momentum in these directions (towards many-core), we expect that these types of systems will become easier to leverage and will, therefore, be more readily adaptable to these and similar astrophysical algorithms.

We note that the publications and software developed for this work are available at our project website: http://lcdm.astro.illinois.edu.

Synergistic Activities

Given our leadership position in this burgeoning field, which was facilitated by this NASA award, we have expended our efforts to both improve the impact of our work and to communicate our results to a wider audience. in this manner, we have vastly increased the impact of this research carried out under this award. First, we have helped organize and run several workshops, symposia, and conferences devoted to the application of novel hardware technology to scientific challenges. Initially we worked with the Reconfigurable Systems Summer Institute (RSSI) in CY 2006, 2007, and 2008, which were all held at NCSA. With these venues, we were able to demonstrate, in person, our algorithms and techniques developed for reconfigurable systems under this award.

After this, we worked with the successor to RSSI, the 2009 Symposium on Application Accelerators in High-Performance Computing (SAAHPC), where we demonstrated additional work on reconfigurable systems and our work leveraging GPUs to an international audience. In addition, we worked with Professor Wen-mei Hwu and David Kirk (Chief Scientist at nVidia) to leverage our algorithmic challenges as student projects for their ECE 498AL Parallel Programming Many-Core systems course. Initially this was offered on the Illinois campus; but subsequently, this course is now offered through the Great Lakes Consortium Virtual Computational School under the NSF funded Blue Waters project.

Finally, in 2009, we held the *Path to Petascale: Adapting GEO/CHEM/ASTRO Applications for Accelerators and Accelerator Clusters* workshop at NCSA. This workshop invited researchers working in these three domain areas to gather and discuss techniques by which this new technology could be leveraged to accelerate scientific discovery. As a result of this successful workshop, we have organized and edited a special edition of Computers in Science and Engineering that will appear soon that is dedicated to scientific applications that are leveraging this new technology.

Personnel:

During the course of this project, a number of personnel were funded. In this section, we list these scientists, and their current institution and position.

Principal Investigator:
Dr. Robert Brunner, Associate Professor of Astronomy, University of Illinois.

Senior Researchers:

Dr. Nicholas Ball, Research Scientist, Herzberg Institute of Astrophysics

Dr. Volodymyr Kindratenko, Senior Research Scientist, National Center for Supercomputing Applications, University of Illinois.

Dr. Adam Myers, Research Scientist, Department of Astronomy, University of Illinois

Dr. Brian Wilhite, Assistant Professor of Physics, Elmhurst College

Graduate Students:

Brett Hayes, Graduate Student, University of Illinois
Dr. Britt Lundgren, Postdoctoral Research, Yale University
Dr. Ashley Ross, Postdoctoral Research, Portsmouth University
Dr. Natalie Strand, Postdoctoral Research, University of Illinois
Yiran Wang, Graduate Student, University of Illinois

Undergraduate Students:

Catherine Grier, Graduate Student, Ohio State University

Publications:

N. Ball and R.J. Brunner, 2010, *Data Mining and Machine Learning in Astronomy*, International Journal of Modern Physics D, in press.

A.J. Ross and R.J. Brunner, 2009, *Halo-model analysis of the clustering of photometrically selected galaxies from SDSS*, MNRAS, 399, 878

B.F. Lundgren et al., 2009, *A Cross-Correlation Analysis of Mg II Absorption Line Systems and Luminous Red Galaxies from the SDSS DR5*, ApJ, 698, 819

V. Kindratenko and R. Brunner, 2009, *Implementation of the two-point angular correlation function on a high-performance reconfigurable computer*, Scientific Programming

A.J. Ross and R.J. Brunner, 2009, *Halo-model analysis of the clustering of photometrically selected galaxies from SDSS*, MNRAS, 399, 878

V. Kindratenko and R. Brunner, 2009, *Accelerating Cosmological Data Analysis with FPGAs*, In Proc. IEEE Symposium on Field-Programmable Custom Computing Machines - FCCM'09

D. Roeh, V. Kindratenko, 2009, R. Brunner, *Accelerating Cosmological Data Analysis with Graphics Processors*, in Proc. 2nd Workshop on General-Purpose Computation on Graphics Processing Units workshop - GPGPU-2

V. Kindratenko, A. Myers, and R. Brunner, 2009, *Implementation of the two-point angular correlation function on a high-performance reconfigurable computer*, Scientific Programming, 17, 3, 247

J. Dolence and R.J. Brunner, 2008, *Fast Two-Point Correlations of Extremely Large Data Sets*, in the Proceedings of the 9th LCI International Conference on High-Performance Clustered Computing.

N.M. Ball, R.J. Brunner, and A.D. Myers, 2008, *Robust Machine Learning Applied to Terascale Astronomical Datasets*, in the Proceedings of the 9th LCI International Conference on High-Performance Clustered Computing.

N.M. Ball et al., 2008, *Galaxy Colour, Morphology, and Environment in the Sloan Digital Sky Survey*, MNRAS, 383, 907

A.J. Ross, R.J. Brunner, and A.D. Myers, 2008, *Normalization of the Matter Power Spectrum via Higher-Order Angular Correlations of Luminous Red Galaxies*, ApJ, 682, 737

A.D. Myers et al., 2008, *Quasar Clustering at 25 h^{-1} kpc from a Complete Sample of Binaries*, ApJ, 678, 635–646

N.M. Ball et al., 2008, *Robust Machine Learning Applied to Astronomical Datasets III: Probabilistic Photometric Redshifts for Galaxies and Quasars in the SDSS and GALEX*, ApJ, 683, 12

N.E. Strand et al., 2008, *AGN Environments in the Sloan Digital Sky Survey I: Dependence on Type, Redshift, and Luminosity*, ApJ, 688, 180

V. Kindratenko, R. Brunner, A. Myers, *Dynamic load-balancing on multi-FPGA systems: a case study*, In Proc. 3rd Annual Reconfigurable Systems Summer Institute - RSSI'07, 2007

R. Brunner, V. Kindratenko, and A. Myers, 2007, *Developing and Deploying Advanced Algorithms to Novel Supercomputing Hardware*, In Proc. NASA Science Technology Conference - NSTC'07

V. Kindratenko, R. Brunner, A. Myers, 2007, *Mitrion-C Application Development on SGI Altix 350/RC100*, In Proc. IEEE Symposium on Field-Programmable Custom Computing Machines - FCCM'07

V. Kindratenko, C. Steffen, R. Brunner, 2007, *Accelerating Scientific Applications with Reconfigurable Computing: Getting Started*, Computing in Science and Engineering, 9, 5, 70

A.D. Myers et al., 2007, *Clustering Analyses of 300,000 Photometrically Classified Quasars. I. Luminosity and Redshift Evolution in Quasar Bias*, ApJ, 658, 85

A.D. Myers et al., 2007, *Clustering Analyses of 300,000 Photometrically Classified Quasars. II. The Excess on Very Small Scales*, ApJ, 658, 99

N.M. Ball et al., 2007, *Robust Machine Learning Applied to Astronomical Datasets II: Quantifying Photometric Redshifts for Quasars Using Instance-Based Learning*, ApJ, 663, 774

A.J. Ross, R.J. Brunner, and A.D. Myers, 2007, *Higher-Order Angular Galaxy Correlations in the SDSS: Redshift and Color Dependence of non-Linear Bias*, ApJ, 665, 67

A.J. Ross, R.J. Brunner, and A.D. Myers, 2006, *Precision Measurements of Higher Order Angular Galaxy Correlations Using 11 Million SDSS Galaxies*, ApJ, 649, 48

N.M. Ball et al., 2006, *Robust Machine Learning Applied to Astronomical Data Sets. I. Star-Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees*, ApJ, 650, 497

Presentations:

R.J. Brunner, 2010 *Grand Questions, Massive Data, Monumental Challenges*, Statistics Department, University of Illinois

V. Kindratenko, R.J. Brunner, G. Shi, D. Roeh, A. Martinez, 2009, *Investigating Application Analysis and Design Methodologies for Computational Accelerators*, NCSA Technical Report

V. Kindratenko, D. Roeh, G. Shi, R. Brunner, 2009, *Accelerating Cosmology Codes*, Path to Petascale: Adapting GEO/CHEM/ASTRO Applications for Accelerators and Accelerator Clusters Workshop

V. Kindratenko, R. Brunner, G. Shi, D. Roeh, A. Martinez, 2009, *Investigating Application Analysis and Design Methodologies for Computational Accelerators*, NCSA Technical Report

N.M. Ball, R.J. Brunner, and A.D. Myers, 2008, *Robust Machine Learning Applied to Terascale Astronomical Datasets*, Astronomical Data Analysis Software and Systems XVII

V. Kindratenko, D. Roeh, 2008, *Internal NCSA GPU Programming Tutorial*

V. Kindratenko, C. Steffen, 2008, *Introduction to Reconfigurable Computing Tutorial*, NCSA

R.J. Brunner, 2008, *Invited Panelist: Great Lakes Consortium for Petascale Computing*, Virtual School of Virtual School of Computational Science & Engineering

T. El-Ghazawi, D. Buell, K. Gaj, V. Kindratenko, 2007, *Reconfigurable Supercomputing Tutorial*, IEEE/ACM Supercomputing

V. Kindratenko, 2007, *High Performance Computing on FPGAs: challenges and opportunities*, Panel on Key Challenges presented by next generation hardware systems, Key Challenges in Modeling and Simulation Fall Creek Falls conference

V. Kindratenko, 2007, *Dynamic Load-Balancing on Multi-FPGA Systems: A Case Study*, Reconfigurable Systems Summer Institute - RSSI'07

V. Kindratenko, 2007, *Accelerating Cosmology Applications: from 80 MFLOPS to 8 GFLOPS in 4 Steps*, SRC's User Meeting

V. Kindratenko, 2007, *Accelerating Scientific Applications with Reconfigurable Computing*, Dept. of Computer and Information Sciences, University of Alabama at Birmingham

V. Kindratenko, 2007, *Developing and Deploying Advanced Algorithms to Novel Supercomputing Hardware*, NASA Science Technology Conference - NSTC'07

V. Kindratenko, 2007, Mitrion-C *Application Development on SGI Altix 350/RC100*, IEEE Symposium on Field-Programmable Custom Computing Machines - FCCM'07

B.P. Hayes, 2007, *Angular Power Spectrum Estimation using High Performance Reconfigurable Computing*, Reconfigurable Systems Summer Institute - RSSI'07

V. Kindratenko, R.J. Brunner, and A.D. Myers, 2007, *Mitrion-C Application Development on SGI Altix 350/RC100*, IEEE Symposium on Field-Programmable Custom Computing Machines

R.J. Brunner, V. Kindratenko, and A.D. Myers, 2007, *Developing and Deploying Advanced Algorithms to Novel Supercomputing Hardware*, The NASA Space Technology Conference

R.J. Brunner, 2006, *Addressing Cosmological Questions by using Reconfigurable Computing,* Reconfigurable Systems Summer Institute

N.M. Ball, R.J. Brunner, A.D. Myers, and D.Tcheng, 2006, *Robust Classification of 143 Million SDSS Objects Via Decision Tree Learning*, The American Astronomical Society Meeting

A. Ross, R.J. Brunner, and A.D. Myers, 2006, *Precision Measurements of Higher-Order Angular Galaxy: Correlations Using 10 Million SDSS Galaxies*, The American Astronomical Society Meeting